AIL 722: Reinforcement Learning

Lecture 40: Bandits, Course Conclusion

Raunak Bhattacharyya

Example: Treating a Broken Toe

- Surgery
- Buddy taping
- Do nothing

- Surgery: $Q(a^1) = \theta_1 = 0.95$
- Buddy taping: $Q(a^2) = \theta_2 = 0.9$
- Do nothing: $Q(a^3) = \theta_3 = 0.1$

- After 6 weeks: do an X-ray to check whether healed (1) or not (0)

Example from Emma Brunskill

Greedy Approach

Take action a^1 ($r \sim$Bernoulli(0.95)), get 0, $\hat{Q}(a^1) = 0$

Take action a^2 ($r \sim$Bernoulli(0.90)), get $+1$, $\hat{Q}(a^2) = 1$

Take action a^3 ($r \sim$Bernoulli(0.1)), get 0, $\hat{Q}(a^3) = 0$

$$a_t^* = \arg\max_{a \in \mathcal{A}} \hat{Q}_t(a)$$

Today’s Outline

- Hoeffding inequality
- Upper confidence bound
- Course conclusion

Correction: Regret

- T denote the total number of rounds or time steps.
- K denote the number of possible actions (arms).
- $a_t \in \{1, \ldots, K\}$ denote the action (arm) chosen by the agent at time t.
- $r_{a_t}(t)$ denote the reward obtained by choosing action a_t at time t.
- a^* denote the optimal action, which is the action with the highest expected reward:

$$a^* = \arg \max_{a \in \{1,\ldots,K\}} \mathbb{E}[r_a]$$

Correction: Regret

The *cumulative regret* $R(T)$ after T rounds

$$R(T) = \sum_{t=1}^{T} (\mathbb{E}[r_{a^*}] - \mathbb{E}[r_{a_t}])$$

- Maximise cumulative reward
    - Equivalent to minimise total regret

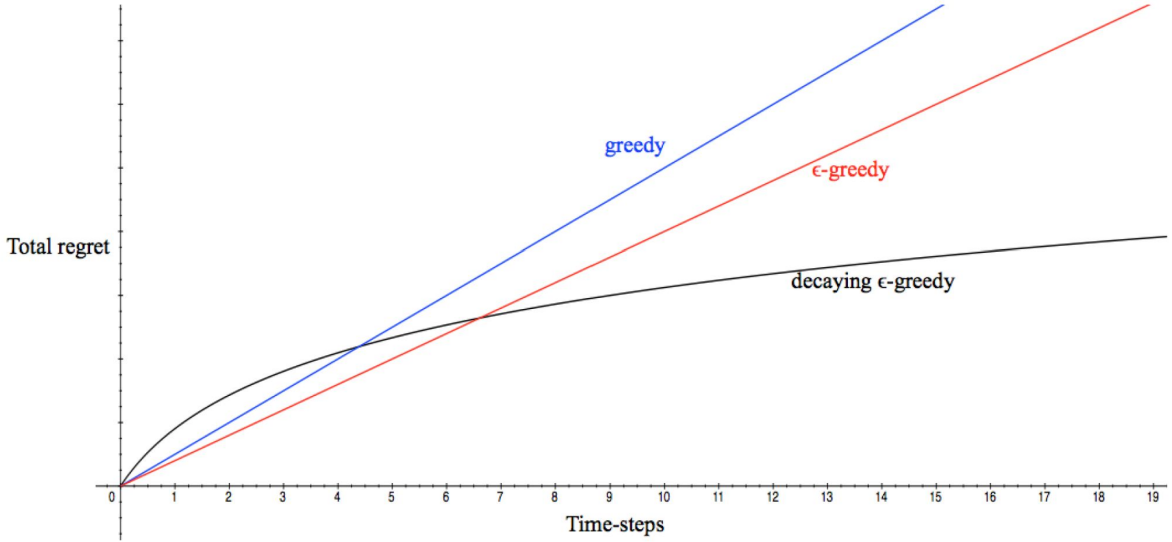
Regret: Greedy Approach

$$R(T) = \sum_{t=1}^{T} (\mathbb{E}[r_{a^*}] - \mathbb{E}[r_{a_t}]) \qquad R(T) = T \cdot \mathbb{E}[r_{a^*}] - \sum_{t=1}^{T} \mathbb{E}[r_{a_t}]$$

Action	Optimal Action	Observed Reward	Regret
a^1	a^1	0	0
a^2	a^1	1	0.05
a^3	a^1	0	0.85
a^2	a^1	1	0.05
a^2	a^1	0	0.05

Cannot evaluate regret in real settings because requires knowledge of the true best action

Regret: Rate of Growth



$$\lim_{t \to \infty} L_t \geq \log t \sum_{a | \Delta_a > 0} \frac{\Delta_a}{D_{KL}(\mathcal{R}^a \| \mathcal{R}^{a^*})}$$

Adapted from Emma Brunskill

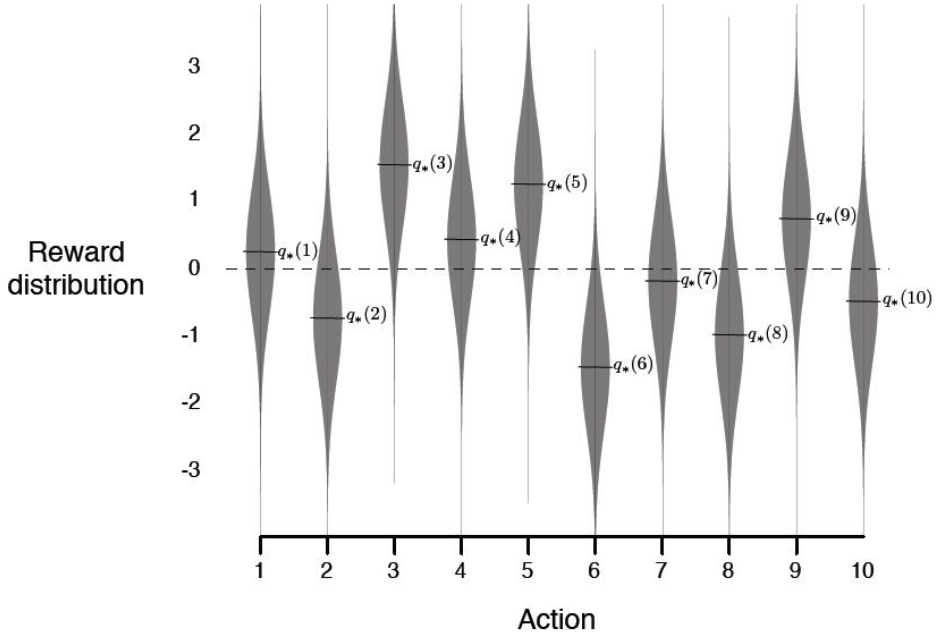
Optimism under Uncertainty

$$A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

$$\lim_{t\to\infty} L_t \leq 8 \log t \sum_{a|\Delta_a>0} \frac{1}{\Delta_a}$$

10-armed Testbed

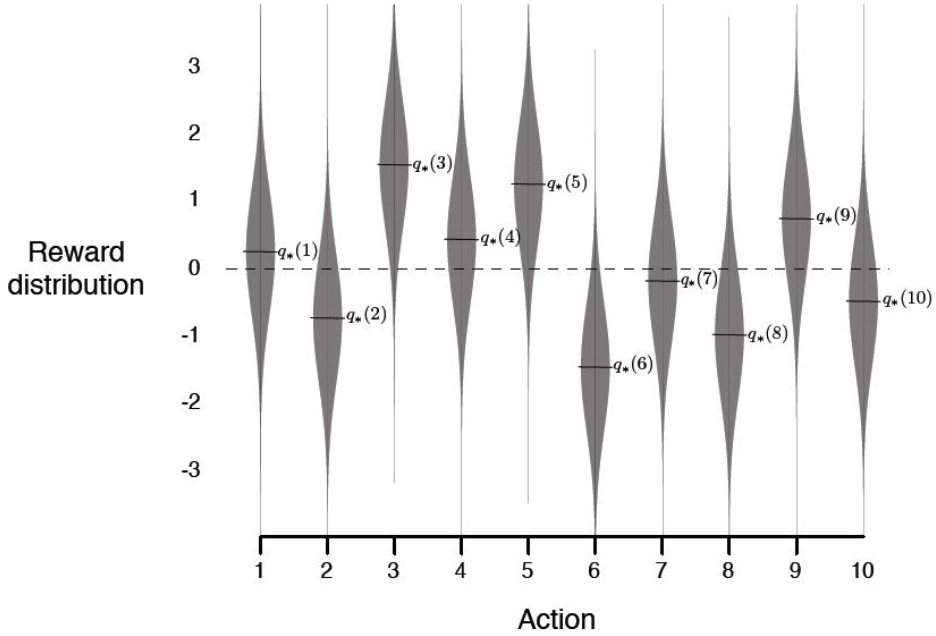
- 2000 10-armed bandit problems
- For each bandit problem, set the true action values by sampling from a standard Normal distribution
- Then, when we test our proposed learning algo, the obtained reward at each timestep is sampled from a Normal(Q(ai),1)



Section 2.3, Reinforcement Learning: An Introduction, Sutton & Barto

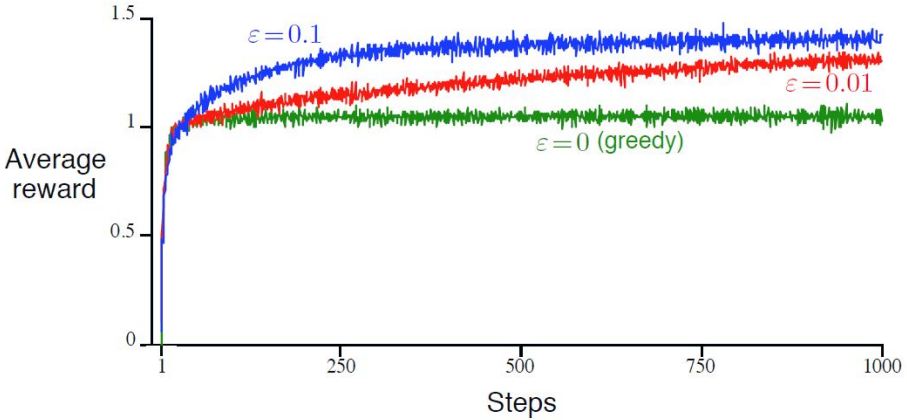
10-armed Testbed

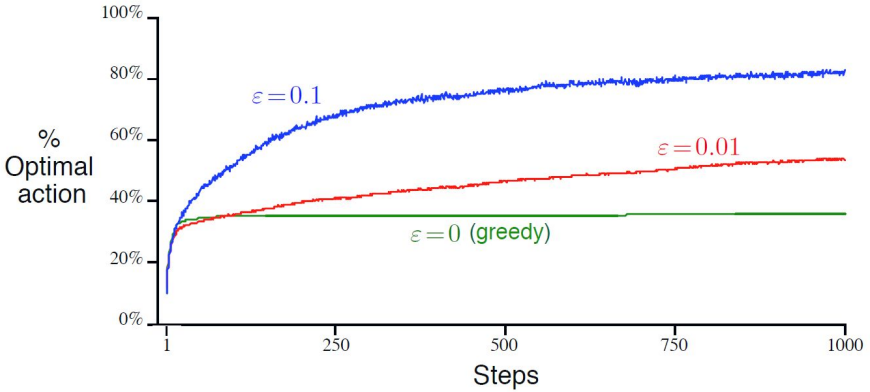
- 2000 10-armed bandit problems
- One run: 1000 timesteps on a bandit problem
- Do 2000 runs, each run for a different bandit problem. This establishes the learning algo performance



Section 2.3, Reinforcement Learning: An Introduction, Sutton & Barto

Performance on 10-armed Testbed: Eps-Greedy



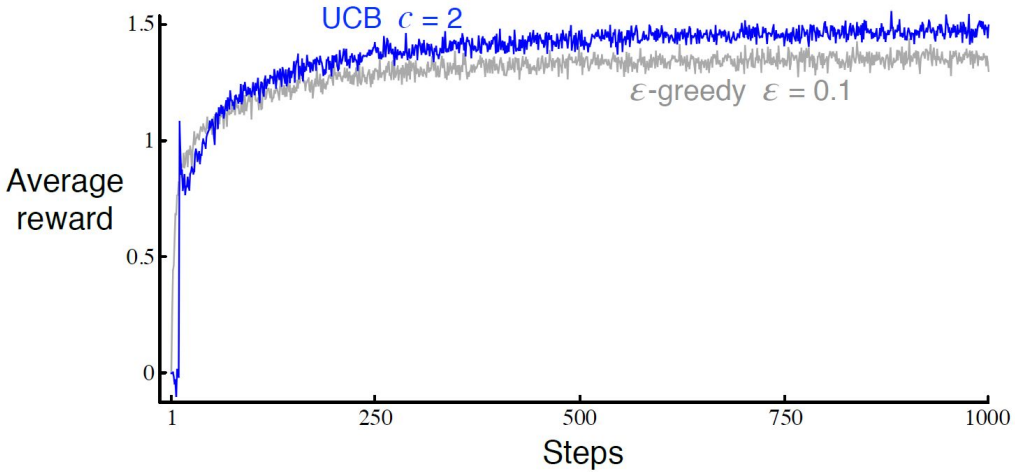


Section 2.3, Reinforcement Learning: An Introduction, Sutton & Barto

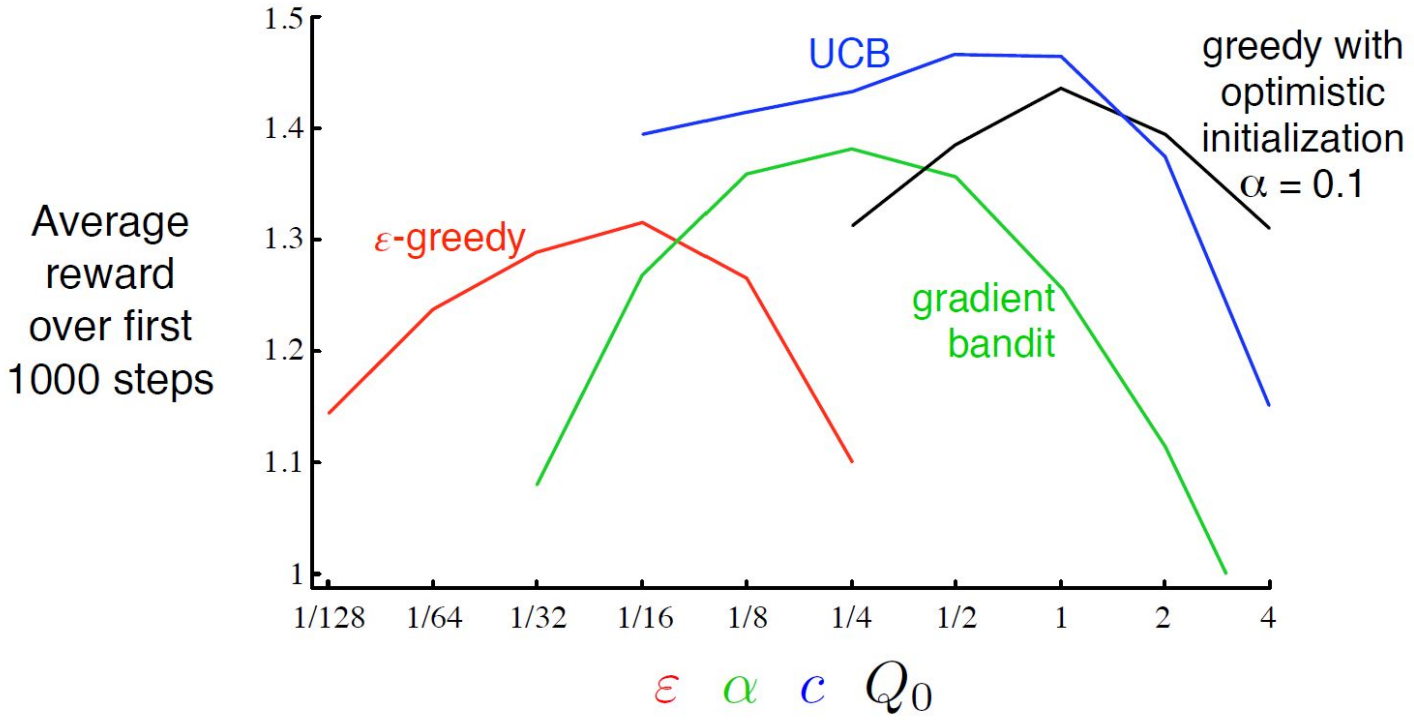
UCB algorithm

$$A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

Reg(T) is O(logT)



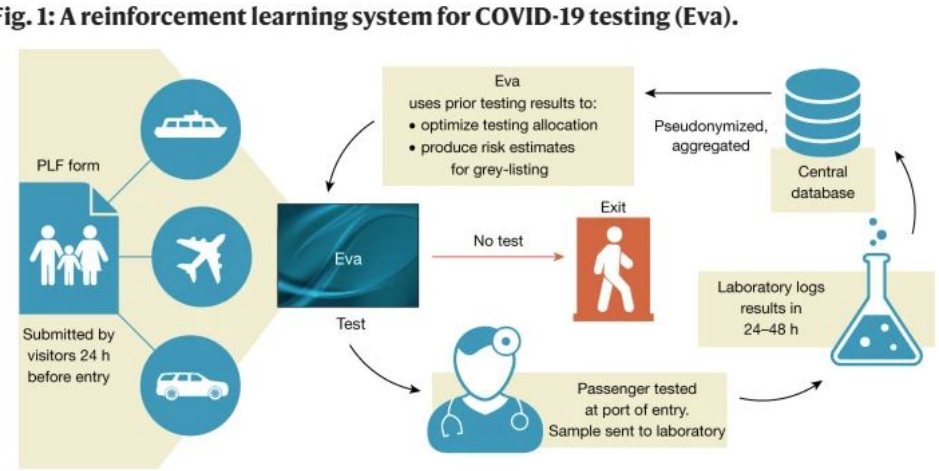
Section 2.7, Reinforcement Learning: An Introduction, Sutton & Barto



Section 2.10, Reinforcement Learning: An Introduction, Sutton & Barto

Applications of Bandits

- Advert placement
- Recommender systems
- Tree search
- A/B testing



Source: Bastani, Nature 2021

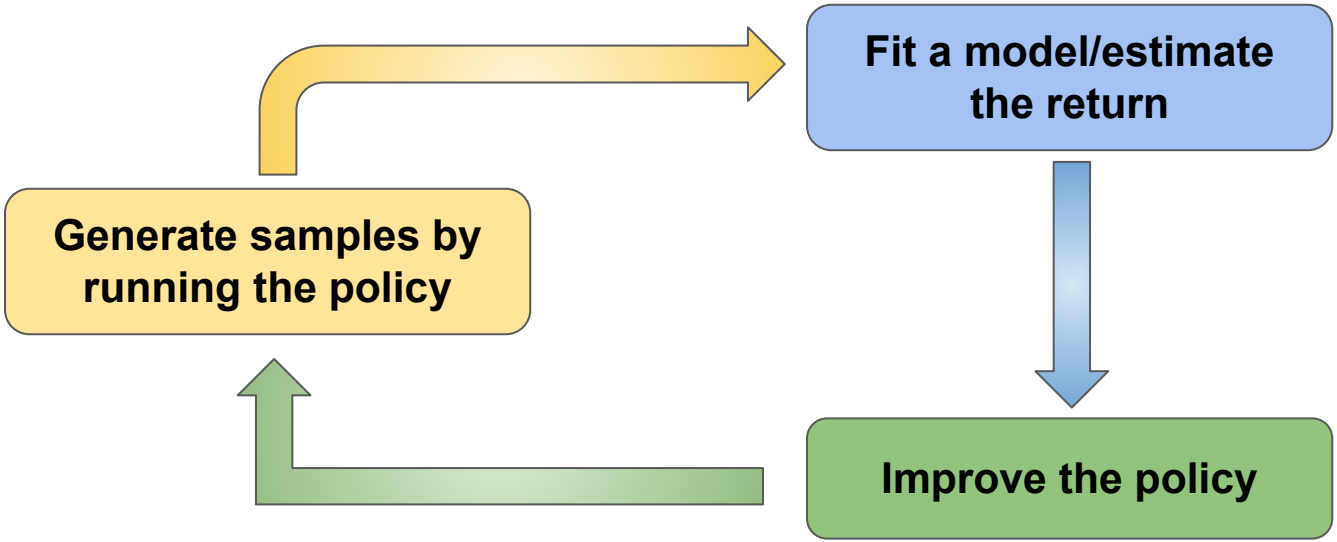
Section 1.2, Bandit Algorithms, Lattimore & Szepesvari

Applications Beyond Robots and Games

- Optimising image generation
- Transportation
- Chip design
- Covid screening

Source: DDPO

Unifying Picture of RL



Adapted from Sergey Levine

Revisiting Algorithm Types

$$\theta^* = \arg\max_\theta \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_t r(s_t, a_t) \right]$$

- Policy gradients
  - Find gradient of objective, then gradient ascent
- Value-based
  - Estimate V or Q (no explicit policy)
- Actor-critic
  - Estimate V or Q of current policy, use to get better gradient estimate

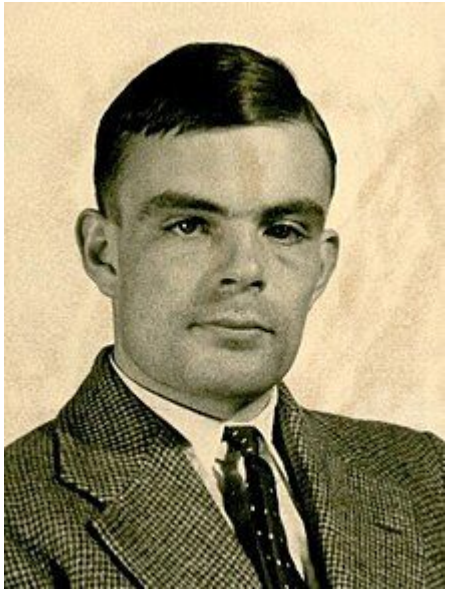
Where Next

- Where do rewards come from?
- Humans reuse past knowledge, policies need to be retrained
- Humans can learn incredibly quickly, RL is slow
- Humans can predict, how should agents use predictions

Conclusion

> INSTEAD OF TRYING TO PRODUCE A PROGRAMME TO SIMULATE THE ADULT MIND, WHY NOT RATHER TRY TO PRODUCE ONE WHICH SIMULATES THE CHILD'S? IF THIS WERE THEN SUBJECTED TO AN APPROPRIATE COURSE OF EDUCATION ONE WOULD OBTAIN THE ADULT BRAIN.

Source: LibQuotes



Alan Turing, Source: Wikipedia